電子掲示板における発言情報を利用したコミュニケーション阻害行為の検出手法の提案(3.2 第5回情報シナジー研究会，3．研究活動報告)

電子掲示板における発言情報を利用した コミュニケーション阻害行為の 検出手法の提案

一藤 裕* 今野 将** 曽根秀昭**
* 東北大学大学院情報科学研究科
** 東北大学情報シナジーセンター

概要　現在、インターネット上の電子掲示板では、コミュニケーションを阻害する発言がたびたび出現している。このような発言が繰り返されると、掲示板は荒れ、利用者の減少など悪影響を及ぼす。そのため、このような発言を発見する手法が必要とされている。本論文では、コミュニケーションを阻害する発言によく出現する単語に着目する。発言単体に出現する単語の数と掲示板全体で出現した数の総数を算出し、出現率から、各単語の重みを設定・評価を行い、コミュニケーション阻害行為の検出を目指す。

A detecting method of BBS vandalism based on information of comments

Yu Ichifuji*, Susumu Konno**, Hideaki Sone**
* Graduate School of Information Sciences, Tohoku University
** Information Synergy Center, Tohoku University

Abstract　Electronic bulletin board systems (BBS) have problems with vandalism. It is necessary for an operator to find such problems quickly when they happen. For detecting such vandalism in BBS, we focus on the words which are used for vandalism. We count the number of such words in each comment and in the BBS. Using two kinds of numbers, we calculate the appearance ratio. The weight of words is determined by such a ratio, and used to detect such vandalism. The results of the detection are shown, and the efficiency of finding such vandalism is discussed.

電子掲示板における発言情報を利用したコミュニケーション阻害行為の検出手法の提案

○一藤 裕†、今野 将‡、曽根 秀昭‡

† 東北大学大学院情報科学研究科
‡ 東北大学情報シナジーセンター

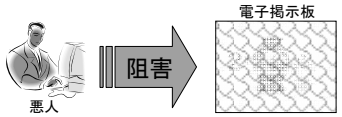
目次

1. 序論
2. 発言者IDに着目した新手法
3. 荒み度RFの評価方法
4. 検証実験
5. 結論

1.1 背景

- 電子掲示板
  - □非リアルタイムの意見・情報交換ツール
  - □教育現場・企業・一般など幅広く利用

正常なコミュニケーションを阻害するユーザが存在



1.1 背景

- コミュニケーション阻害行為の種類（荒らし行為）
  1. コミュニケーションを阻害する書き込み
  2. 他者によるプライバシー情報の公開
  3. 閲覧者を不快にさせる書き込み
  4. 他者を挑発し、反応を楽しむ書き込み(煽り・釣り)
  5. 無用の書き込みを連続で行い正常な閲覧を阻害する
  6. 根拠のない情報を公開し、対象を貶める

管理者は荒らし行為の発生を把握する必要がある

1.2 既存サービス

- 有人監視サービス（ピットクルー株式会社）
  - □NGワードを登録
  - □24時間NGワードが出現するたびに人間が確認

•利点
文脈を理解し判断を下せるため、確実に発見可能
•欠点
人件費が高く、個人運営の掲示板向きではない

人間を使わずに自動化できれば・・・

1.3 前手法の紹介

- 荒らし行為発見自動化のためのアプローチ
  - □ 文脈理解
  - □ NGワードフィルタ
  - □ 発言間の係り受け

  ⇒ 単純な方法で荒らし行為発見を目指す

  - □ 掲示板の有向グラフ化
- 対象とする荒らし行為
  - コミュニケーションを阻害する書き込み
  - 閲覧者を不快にさせる書き込み
  - 他者を挑発し、反応を楽しむ書き込み(煽り・釣り)

1.3 前手法の紹介

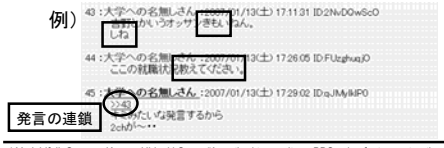
- 対象掲示板の形式
  - □発言は発言した順番に表記
  - □特定の発言に対し発言する場合、アンカーと呼ばれる記号を利用

例）

1.3 前手法の紹介

- 荒み度を用いた監視支援手法*
  - □着目点
    - **心理的影響を与える単語**
    - **発言の連鎖**
    - **掲示板の雰囲気（荒み度）**

例）



*:Yu Ichifuji, Susumu Konno, Hideaki Sone, "A method to monitor a BBS using feature extraction of text data", International Conference on Human.Society@Internet, (2005) 349-352

1.3 前手法の紹介

- 荒み度を用いた監視支援手法

  荒み度(Ruination Figure [RF])とは
  掲示板の雰囲気を数値化した指標
  - □好感・嫌悪感を与える単語
  - □発言の連鎖

単語による影響力 ＋ 発言の連鎖による影響力 ＝ 発言の影響力

$\sum$ 発言の影響力 ＝ 荒み度 RF

1.3 前手法の紹介

- 荒み度を算出するための定義

pw ：相手に好感を与える単語の集合(辞書)

nw ：相手に嫌悪感を与える単語の集合(辞書)

pww ：pw登録単語が持つ重み(tf-idf法で算出)

nww ：nw登録単語が持つ重み(tf-idf法で算出)

cn ：一つの発言において各単語が出現した回数

- Ws (Word score) 単語による発言の影響力

$$Ws(t)=\sum_{i}^{pw} pww(i)\cdot cn(i)+\sum_{j}^{nw} nww(j)\cdot cn(j)$$

1.3 前手法の紹介

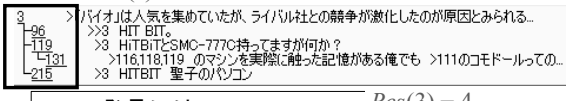
- ccs (Comment Chain Score)
  発言の連鎖数から算出する発言の影響力

$$ccs(t)=i\log\{Res(t)\}$$

$i=\pm 1: Ws(t)$に依存

t：発言ナンバー

$Res(t)$：発言 t の連鎖数



3 の発言には4つのレス	$Res(3)=4$
119の発言には1つのレス	$Res(119)=1$

1.3 前手法の紹介

- Ss (Statement Score)
  各発言が与える影響力

$$Ss(t)=Ws(t)+\frac{\{ccs(t)\cdot Max(|Ws|)\}}{Max(|ccs|)}$$

$i=\pm 1: Ws(t)$に依存

t：発言ナンバー

- RF (Ruination Figure)
  掲示板の評価指標

$$RF(t)=\sum_{1}^{t} Ss(j)$$

1.3 前手法の紹介

- 前手法の問題点
  - 荒らし行為すべてを発見できない（手法が単純すぎるため）
    - コミュニケーションを阻害する書き込み
    - 閲覧者を不快にさせる書き込み
    - 他者を挑発し、反応を楽しむ書き込み(煽り・釣り)

荒らし行為を特徴で分類し、それぞれの特徴にあった荒らし行為を発見する手法を確立する。最終的に複数の手法の組み合わせで様々な荒らし行為の発見を実現する

1.4 本発表の目的

- 同一人物が複数回関与する荒らし行為発見のための手法を提案する

同一人物が複数回出現する荒らし行為の例

- 釣り・煽り
  - 故意に炎上（フレーム）を引き起こす行為
- 叩き
  - **煽り**や荒らしに反応し、相手を非難・糾弾・指弾する行為
- 罵り合い（自作自演も含む）
  - 同一人物が複数の人間が行っているように見せる行為
  - 数人が言い争いを行い他の閲覧者を不快にさせる行為

1.5 各荒らし行為の定義

- 釣り・煽り・叩きの例

釣り

14：大学への名無しさん：2007/01/26(金) 00:04:51 ID:VYkwq00h0
理Ⅲから受ければ理Ⅱはゴミカスか。

叩き

15：大学への名無しさん：2007/01/26(金) 00:08:42 ID:3YcmAR8h0
>>13 市ね。医学部が偏差値のみ事実だが
変人ばっかりにかねがけて、国の税金をろうびさせてうえに、
金ばっかり追求するから、研究レベルも世界の中で
低レベル。

煽り or 叩き

16：大学への名無しさん：2007/01/26(金) 00:10:08 ID:Fb15sEPD0
>>14
偏差値が低いとはいえども東大は偏れない
確かに大半は大したことないが、医者になりたくないから理12逝く奴がいる
ただし、東北の歯医はマジでカス。例外無く東大未満のバカしかいない。

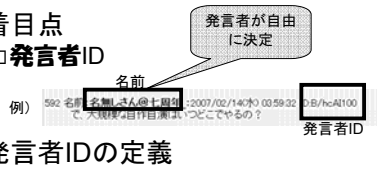
煽り or 叩き

17：大学への名無しさん：2007/01/26(金) 00:12:01 ID:VYkwq00h0
>>16
腹目でどんなに頑ろうと、キミはもう一生東北大生にはなれないんだよ。
毎年この時期はつらいなー。w
辛い受験の思い出がよみがえっちゃうもんなー。w

2章 発言者IDに着目した新手法

2.1 荒らし行為発見の準備

- 着目点
  - □**発言者**ID



- 発言者IDの定義
  - □IPアドレス（書き込む掲示板や日付）により作成
  - □IDは個別に付与
  - □同一ID＝同一人物とみなす

複数回発言しているIDのSsを強調すれば
対象荒らし行為の発見がより容易になるのでは？

2.1 荒らし行為発見の準備

- 複数回同一IDが関与した荒らし行為の特徴・分類
  - □ CASE1
    - ある発言を行い、一斉に反論を受け、それを認めず応戦し不快な単語を伴い荒れる。返答の仕方・発言者によってさらに分類
  - □ CASE2
    - 釣りを行い、予想範囲内の反応を得て荒れるor叩きが発生し荒れる
  - □ CASE3
    - 一人の人物が脈絡なく複数の発言に対し文句をつける発言を連投する。

2.1 荒らし行為発見の準備

- CASE1
  - □1対多数
  - □噛み付かれた人間が再度出現する可能性大
    噛み付いた人間が再度出現する可能性も
  - □返答の仕方が2パターン
    - パターン1：一つの発言で全てに返答
    - パターン2：一つの発言で一つの返答。
      連続発言になる

2.1 荒らし行為発見の準備

- CASE2
  - □釣った人間は反応した人間に対して罵倒するため、同一人物が再度出現する
- CASE3
  - □連続発言のため、同一人物が繰り返し出現する

CASE1,2,3を発言IDに着目し、新たな算出方法を提案することにより荒らし行為の発見を目指す

2.2 提案手法の算出方法

- 算出手順
  1. ID別に発言回数を算出
  2. N回以上かつN回の発言のSsの総和が負の値のみT倍し、Ss'とする

     Ss(t) ＝ 単語による影響力 ＋ 発言の連鎖による影響力
  3. Ss'を用いてRFを算出し、グラフ出力し目視にて評価する

発言回数がN回以上かつSsの総和が負のとき

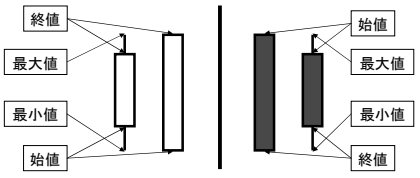
$$Ss(t)' = T * Ss(t)$$

3章 荒み度RFの評価方法

3. 荒み度の評価方法

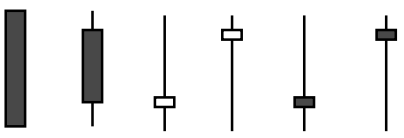
- RFの表現方法
  - □RFを10区切りに分ける
    (RF(1)~RF(10)、RF(11)~RF(20)、・・・)
  - □1区切りごとに最大値、最小値、始値、終値を抽出しローソク足(株価の表現で使われる)で表現

例）



3. 荒み度の評価方法

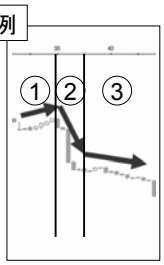
- 着目すべきローソク足



これらのアイテムが他に比べ長く、密集している範囲が荒れている疑いのある範囲

3. 荒み度の評価方法

例



- 荒らし行為の発見
  - □荒み度の変化の度合いから抽出
    - ①では、緩やかに上昇
    - ②で、急激に下降
    - ③では、緩やかに下降
  - □②で話題の転換が発生
    急激に下降したため
    「荒らし行為が発生したのでは？」
    と相対的に判断することとする

4章 検証実験

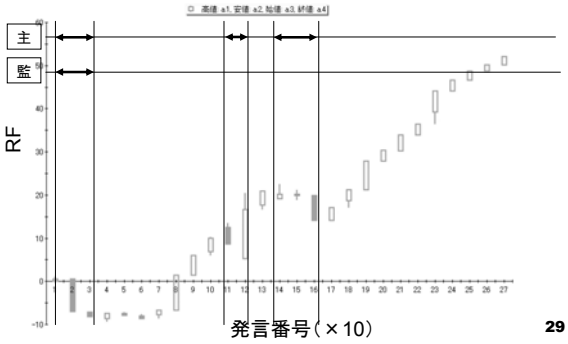
4.1 実験対象

- 対象掲示板とその評価
  - □“2ちゃんねる”の大学受験掲示板に限定
  - □主観評価
    - 実際に読み、荒れている範囲を抽出
    - 4人の学生による判断
- 検証方法
  - □前手法と提案手法のN,Tを変化させた場合との比較

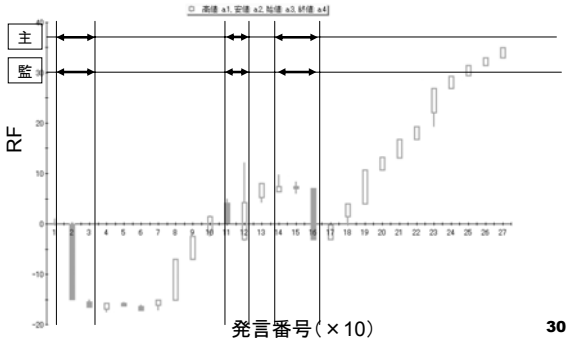
4.2 評価事例

- 例示掲示板
  - □“東北大学 理系専用 part2”
- 荒らし行為の範囲
  - □5-22　同一IDが出現する煽り・叩きが存在
  - □110-122　個別による煽り・叩きと通常
  - □131-152　同一IDが出現する煽り・叩きと通常

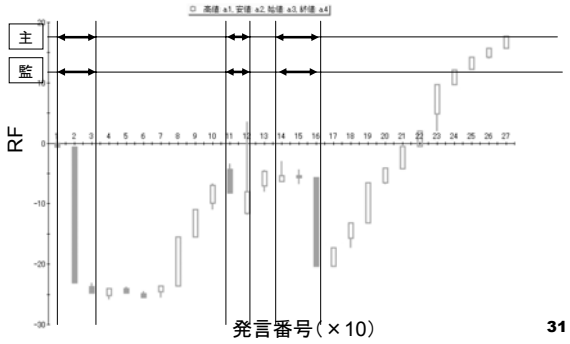
4.3.1 前手法による出力結果



4.3.2 新手法の出力結果（T=2,N=3）



4.3.3 新手法の出力結果（T=3,N=3）



4.4 実験結果

- 前手法では曖昧さのあるグラフに対し、提案手法でははっきりとした差別化ができた

5章 結論

5.1 まとめ

- 対象とするコミュニケーション阻害行為を限定
- 釣り・煽り・叩き・自作自演を発言IDに着目した監視支援手法を提案
- 検証実験により、発言IDに着目することは荒らし行為の範囲を強調できることを示した

5.2 今後の課題

- 荒らし行為には様々なタイプがあるため、複数回登場する人物による煽り・釣り・叩き・自作自演以外にも対応させる必要がある
- T,Nの設定の自動化・閾値の設定